Requested Patent: WO0033215A1

Title:

TERM-LENGTH TERM-FREQUENCY METHOD FOR MEASURING DOCUMENT SIMILARITY AND CLASSIFYING TEXT ;

Abstracted Patent: WO0033215 ;

Publication Date: 2000-06-08 ;

Inventor(s): KANTROWITZ MARK (US) ;

Applicant(s):

KANTROWITZ MARK (US); JUSTSYSTEM PITTSBURGH RESEARCH (US) ;

Application Number: WO1999US25686 19991101 ;

Priority Number(s): US19980201569 19981130 ;

IPC Classification: G06F17/30 ;

Equivalents: AU1907300 ;

ABSTRACT:

A computer implemented method of extracting characterizing terms from a document comprising the steps of extracting terms from the document, counting the number of occurrences of each term extracted from the document to establish a frequency value for each term, counting the characters or strokes in each term to establish a character count for each term, multiplying the frequency value for each term or a monotonic function thereof by the character count for each term for each term to establish a modulus for each term and sorting the terms according to their moduli whereby the terms with the greatest moduli may be accepted as keyword's characteristic of the document's content.

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G06F 17/30	A1	(11) International Publication Number: WO 00/33215
		(43) International Publication Date: 8 June 2000 (08.06.00)

(21) International Application Number: PCT/US99/25686

(22) International Filing Date: 1 November 1999 (01.11.99)

(30) Priority Data:
09/201,569 30 November 1998 (30.11.98) US

(71) Applicant *(for all designated States except US)*: JUST-SYSTEM PITTSBURGH RESEARCH CENTER, INC. [US/US]; 4616 Henry Street, Pittsburgh, PA 15213 (US).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: KANTROWITZ, Mark [US/US]; 5503 Covode Street, Pittsburgh, PA 15217 (US).

(74) Agents: BYRNE, Richard, L. et al.; Webb Ziesenheim Logsdon Orkin & Hanson, P.C., 700 Koppers Building, 436 Seventh Avenue, Pittsburgh, PA 15219-1818 (US).

(81) Designated States: AE, AL, AM, AT, AT (Utility model), AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, CZ (Utility model), DE, DE (Utility model), DK, DK (Utility model), DM, EE, EE (Utility model), ES, FI, FI (Utility model), GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SK (Utility model), SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: TERM-LENGTH TERM-FREQUENCY METHOD FOR MEASURING DOCUMENT SIMILARITY AND CLASSIFYING TEXT

(57) Abstract

A computer implemented method of extracting characterizing terms from a document comprising the steps of extracting terms from the document, counting the number of occurrences of each term extracted from the document to establish a frequency value for each term, counting the characters or strokes in each term to establish a character count for each term, multiplying the frequency value for each term or a monotonic function thereof by the character count for each term for each term to establish a modulus for each term and sorting the terms according to their moduli whereby the terms with the greatest moduli may be accepted as keyword's characteristic of the document's content.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

TERM-LENGTH TERM-FREQUENCY METHOD FOR MEASURING DOCUMENT SIMILARITY AND CLASSIFYING TEXT

BACKGROUND OF THE INVENTION

Information retrieval, text classification, information filtering, text summarization, text keyword extraction and highlighting, document routing and related processes are increasingly needed with the advent of the World Wide Web. Most such methods involve forming vectors of term scores or weights for a document or term, where a term may be a word character, n-gram or phrase. The task is to then find the pairs of vectors that are closest using some definition of closeness, such as the least-squares method or the cosine distance (dot product) method.

The most common methods of computing scores for terms involve statistical information about the document itself, as well as information about the universe of documents that include the document. For example, term frequency multiplied by inverse document frequency (TFIDF) computes the frequency of each term in the document or query and multiplies it by the reciprocal of the frequency of the term across documents (a measure of the rarity of the term). Unfortunately, one must maintain statistics about the set of documents and maintain substantial information external to the document itself. Updating this information is an important problem for such methods.

U.S. Patent No. 4,843,389 entitled "Text Compression and Expansion Method and Apparatus" uses the product of the word length in characters by the word's frequency of occurrence for text compression but not information retrieval. Also, this patent involved the frequency of occurrence of the word in the language in which it is used (i.e., "general usage over a sample of texts from the user's environment"), not the frequency of occurrence of the term in the document being compressed.

U.S. Patent No. 5,182,708 entitled "Method and Apparatus for Classifying Text" focuses on using a document readability metric to distinguish texts in computer manuals from text written by foreigners and native English speakers. It multiplies the quantity log (N/L) / [log(N) -

1], where N is the number of words in the document and L is the number of different words by the correlation coefficient between the word length and the logarithm scaled rank order of word frequency. The latter is evaluated for the particular document, not all documents, and yields a measure of the degree to which polysyllabic words are used by the document. N/L is the average term frequency for the document.

U.S. Patent No. 5,293,552 entitled "Method for Storing Bibliometric Information on Items From a Finite Source of Text, and in Particular Document Postings for Use in a Full-Text Document Retrieval System" makes use of a postulated rank-occurrence frequency relation. It was found that the resulting computed frequencies are too high for high frequency terms and too low for low frequency terms and that taking the square root of the estimated occurrence frequencies yields better results than using the raw occurrence frequencies themselves. The patent concerns the use of this estimation technique to reduce the size of the indexes in the information retrieval algorithm.

The computer method, according to the present invention, does not require any information outside the document being scored and is easy to implement. It is so simple that it would not be expected to work well, but in fact outperforms some existing methods. A document summarizer based on this method is easy to implement and use and requires less memory than other methods. The present invention is also scalable because it does not rely on information outside the document itself and so does not consume more resources as the number of documents increases. Avoiding the need to update this information makes the present invention more scalable than state-of-the-art information retrieval algorthims, making it also highly suitable for distributed information retrieval applications.

The present invention is directed to information retrieval, text filtering, text summarization, text

classification, keyword extraction and related tasks, but not text compression.

SUMMARY OF THE INVENTION

Briefly, according to this invention, there is provided a method for identifying the most descriptive words in a computer text file or stream. The method of extracting characterizing terms from a document comprises a) extracting terms from the document, b) counting the number of occurrences of each term extracted from the document to establish a frequency value for each term, c) counting the characters or strokes in each term to establish a character count for each term, d) multiplying the frequency value for each term or monotonic function of the frequency value by the character count or for each term or a monotonic function of each count to establish a modulus for each term and e) sorting the terms according to their moduli whereby the terms with the greatest moduli may be accepted as characteristic of the document's content. In non-Roman alphabets such as used with Asian languages, the complexity of a word is reflected more by the number of strokes in the word than the number of characters.

By computer text file, it is meant an ASCII or other standard text file comprised of bytes representative of alphanumeric characters grouped together to represent words. By text stream, it is meant a series of bytes representative of alphanumeric characters being grouped together to represent words being transferred serially or in parallel into the computer from a file, keyboard, modem or some other device.

The computer implemented method disclosed herein multiplies the frequency of terms (words or phrases) appearing in a text file or stream (hereafter "document") by the length of each term to establish a modulus (score or weight) by which each term can be ranked. Terms with the highest modulus are the most descriptive of the content or information found in the document. Variations on this method include multiplying term frequency by the logarithm

of term length, multiplying term frequency by the square root of term length, using a stop-list of articles, prepositions and other common terms to eliminate such terms from the calculation and stemming or truncating words to standard word form. Additionally, the resulting term weights can be normalized using a fixed table of term constants. These constants would be based on a large training corpus of text documents, but this corpus would not necessarily be the same as the collection of documents being indexed. The idea is not to reintroduce term frequency inverse document frequency (TFIDF) into the formula, but to normalize the term length form frequency (TLTF) values by the typical values for the term so that TLTF then highlights departures from the norm. The possible constants include normalizing term frequencies by the overall frequency of occurrence of the term in the language and normalizing term frequencies by precomputed term rarity values (i.e., multiplying by inverted document frequencies for a reference corpus).

Both term length (TL) and term frequency (TF) are available from the document itself without requiring any external resources. For extracting significant keywords, one presents the n terms (for some number n) with the highest moduli (scores). For summarizing documents, one may present the sentences with the highest total scores. For document similarity, one uses the scores as the elements of the term vector.

The method disclosed herein has application for term weighting and is applicable to information retrieval applications, such as document retrieval; cross-language information retrieval; keyword extraction; document routing; classification; categorization; clustering; document filtering; query expansion; chapter, paragraph and sentence segmentation; spelling correction (i.e., ranking candidate corrections); term, query and document similarity metrics; and text summarization.

BRIEF DESCRIPTION OF THE DRAWING

Further features and other objects and advantages will become clear from th following detailed description made with reference to the drawing which is a flow diagram for explaining a computer program for implementing the invention.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

This is a computer program implemented method. It has been implemented in the Perl language which is particularly useful for processing documents. The Perl language is described in Learning Perl by Randal L. Schwartz and Tom Christiansen (O'Reilly & Associates, Inc. 1997). The main Perl program instructions comprise:

```
&load_file("file");
&gather_freq();
&process_freq();
foreach $key (@result) print " $key";
```

As those skilled in the art know, most programs can be written in any one of a number of languages and using different approaches can cause the computer to follow the same basic steps. The drawing schematically illustrates the main portion of the program.

Referring to the drawing, the first step 10 is to capture the document content line by line. Perl has a standard function for inputting a line of text from a file. The line is stored in a scalar variable. The next step 20 is to isolate each word in the line. Perl has a function for splitting a line of text into an array of words. The array of words is then processed to build a word frequency hash using each new word as a key at 30. A hash is like an array in that it is a collection of scalar values with individual elements indexed by keys associated with each value. Hence, at 30, each word is individually compared to the keys already in the frequency hash and if not present, added to the hash as a key and if present, the value (number) associated with the key is incremented. After each line is processed at 30, a test is made at 40 to

determine if another line of text is available in the document. Steps 10, 20 and 30 are repeated until all lines in the entire document are processed.

In the Perl embodiment, steps 10, 20 and 30 are implemented by the &load_file and &gather_freq subroutines. Portions of the &gather_freq are set forth here.

```
sub gather_freq {
    foreach $line (@file) {
        if ($line =~ /^\</) {
        } else {
        @words = split(/\s+/,$line);
        foreach $word (@words) {
            &add_keyword ("$word");
            } } } }
```

The "&add_keyword" subroutine adds or increments the count of words to the frequency hash "%keyfreq".

The processing step 30 can be made more sophisticated by ignoring certain extremely common words ("stop words"), such as articles and prepositions that have little value in characterizing the content of the document. As each word in the array is encountered, it is compared for membership in a stop word array and if present, it is skipped. Alternatively, the stop words can be eliminated at the end of the process rather than at this time. The processing step 30 can be made even more sophisticated by stemming words to the same form. Hence, the plural form of words can be changed to singular and the past tenses and past participles of verb forms can be truncated to the present tense.

Where the method of this invention is simply to identify the most characteristic terms in a document, the normalizing step 40 need not be implemented. However, if the word characterizing vector created by the method is to be compared with the word characterizing vectors of other documents, normalization is desirable. The simplest normalization technique is to divide all frequency values in the word frequency hash by a constant indicative of the

length of the document. A more complicated normalization process would be to use a table of normalization constants which hold a normalization constant for each word. The normalization process would then divide the frequency value associated with each word key with a constant that either increases or decreases the value according to a preconceived understanding of the word for characterizing the text. It would not be unlikely that both of these normalization techniques be used. Yet another normalization method comprises subtracting the average frequency for the term from the term's frequency and dividing the result by the standard deviation for the frequencies. The average frequency for the term and the standard deviation is computed using a collection of documents. Since the average and standard deviation do not change much when documents are added to the collection, we can consider them to be fixed. This normalization method scales the moduli into a standard unit's domain, identifying the number of standard deviations a way the term's frequency is from the typical frequency for the term. This makes the scaled values comparable between documents and terms.

The next step 60 is to form a keyword array by extracting the keys from the word frequency hash. The keyword array is then sorted by the product of the frequency value for that word multiplied by the length of the word.

```
sub process_freq {
    @result = sort by_freq (keys(%keyfreq));}
sub by_freq {
    $x = $keyfreq {"$a"} * length ("$a");
    $y = $keyfreq {"$b"} * length ("$b");
    if ($x < $y) {
        1;
    } elesif ($x == $y) {
        0;
    }elseif ($x > $y) {
        -1
    }
}
```

At 70, the words at the front of the keyword array (those with the greatest product) are then displayed as the words that best characterize the document. For example, the entire document might be displayed highlighting the five words at the front of the keyword array.

In an extension of this method, the keyword array might be converted to keyword value hash. The value for each keyword may simply be one or it might be the product frequency times word length. The keyword product hashes for two documents can then be combined by dot product multiplication followed by summing the product components to get a factor indicative of the similarity of the content of the two documents.

Having thus described my invention with the detail and particularity required by the Patent Laws, what is desired protected by Letters Patent is set forth in the following claims.

I CLAIM:

1. A computer implemented method of extracting characterizing terms from a document comprising the steps of:

a) extracting terms from the document;

b) counting the number of occurrences of each term extracted from the document to establish a frequency value for each term;

c) counting the characters or strokes in each term to establish a character count for each term;

d) multiplying the frequency value for each term or a monotonic function thereof by the character count for each term or a monotonic function thereof to establish a modulus for each term; and

e) sorting the terms according to their moduli whereby the terms with the greatest moduli may be accepted as keyword's characteristic of the document's content.

2. The method according to claim 1, wherein a step is provided for discarding stop words from the extracted terms.

3. The method according to claim 1, wherein a step is provided for the term counts to be normalized.

4. The method according to claim 3, wherein a step is provided for the term count to be normalized by a constant indicative of the length of the document.

5. The method according to claim 3 or 4, wherein a step is provided for the term count to be normalized by dividing the term count for individual terms by constants for terms that are preconceived to increase or decrease the frequency value according to the import of the term for characterizing the content of the document.

6. The method according to claim 3, wherein a step is provided for the term count to be normalized by subtracting a term specific constant and dividing by another term specific constant.

7. The method according to claim 1, wherein a step is provided for stemming terms extracted from the document to a standard word form.

8. The method according to claim 1 further comprising the steps of counting the keywords in each sentence of the document and displaying the sentence with the most keywords therein.

9. The method according to claim 1 further comprising the steps of generating a score for each sentence in the document based upon the moduli of the words in the sentence and displaying the sentences with the highest scores as a summary of the document.

10. The method according to claim 9, wherein the scores are normalized by the sentence length.

11. The method according to claim 9, wherein the scores are computed by adding the top n moduli for each sentence.

12. The method according to claim 1 further comprising the steps of displaying the entire document with the keywords highlighted.

13. A computer implemented method of comparing the content of two documents comprising extracting characterizing terms from each document by the steps of:

a) extracting terms from the document;

b) counting the number of occurrences of each term extracted from the document to establish a frequency value for each term;

c) counting the characters in each term to establish a character count for each term;

d) multiplying the frequency value for each term or a monotonic function thereof by the character count for each term or a monotonic function thereof to establish a modulus for each term;

e) sorting the terms according to their moduli whereby the terms with the greatest moduli may be accepted as characteristic of each document's content;

f) forming a vector for each document based on a number of terms having the greatest moduli; and

g) generating a factor indicative of the similarity of the two documents.

14. The method according to claim 13, wherein the step for generating a factor indicative of the similarity of the two documents includes forming the dot product of the two vectors.

15. A computer implemented method of document retrieval based on a query including query terms comprising extracting characterizing terms from each document selected by the query comprising the steps of:

a) extracting terms from the document;

b) counting the number of occurrences of each term extracted from the document to establish a frequency value for each term;

c) counting the characters in each term to establish a character count for each term;

d) multiplying the frequency value for each term or a monotonic function thereof by the character count for each term or a monotonic function thereof to establish a modulus for each term;

e) sorting the terms according to their moduli whereby the terms with the greatest moduli may be accepted as characteristic of each document's content;

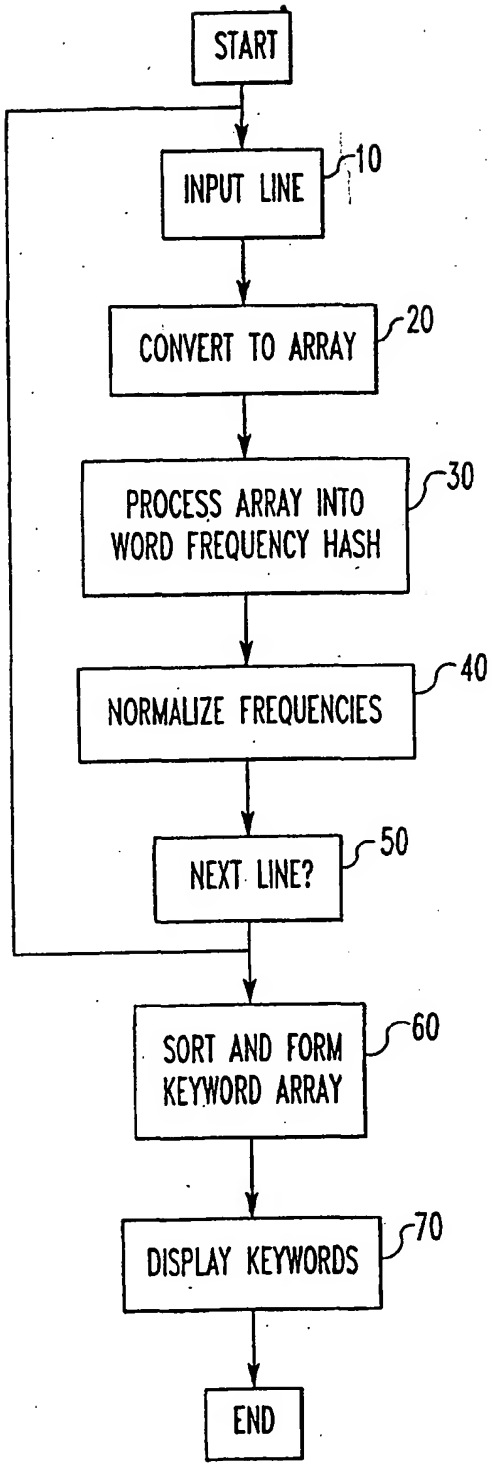
f) forming a vector for each document based on a number of terms having the greatest moduli; and

g) generating a factor indicative of the similarity of the document to the query terms.

16. The method according to claim 5, wherein the step for generating a factor indicative of the similarity of the document to the query terms includes forming the dot product of the document vector and a vector of query terms.

17. The method according to claim 1 comprising the further steps of calling a spell checking program which provides a candidate correction list of terms for each misspelled term encountered and sorting the candidate correction list in ascending order according to modulus.

18. The method according to claim 13, wherein terms in the candidate correction list that do not have a modulus are sorted in ascending order by term length.


START
INPUT LINE
10
CONVERT TO ARRAY
20
PROCESS ARRAY INTO WORD FREQUENCY HASH
30
NORMALIZE FREQUENCIES
40
NEXT LINE?
50
SORT AND FORM KEYWORD ARRAY
60
DISPLAY KEYWORDS
70
END

INTERNATIONAL SEARCH REPORT

Interi. .onal application No.
PCT/US99/25686

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :G06F 17/30
US CL :707/3, 5

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 707/3, 5

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

none

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WEST, EAST

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5,748,953 A (MITZUTANI et al) 05 May 1998, Abstract, lines 1-18.	1-18
X, Y	US 5,544,049 A (HENDERSON et al) 06 August 1996, Abstract, lines 1-17.	1-18
X, Y	US 5,642,502 A (DRISCOLL) 24 June 1997, Abstract, lines 1-27.	1-18

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search 11 FEBRUARY 2000	Date of mailing of the international search report 28 FEB 2000
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer Cheryl Lewis for Virginia Logan Telephone No. (703) 305-8750

Form PCT/ISA/210 (second sheet)(July 1992)*